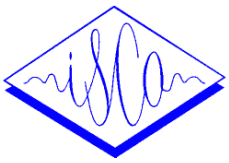

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

THE DEVELOPMENT AND PERCEPTIVE EVALUATION OF A MODEL FOR PARAGRAPH INTONATION IN DUTCH[1]

A. Sluijter[+] *and J.M.B. Terken*[*]

[+]Dept. Linguistics/Phonetics Laboratory Leyden University
P.O. Box 9515, 2300 RA Leiden, The Netherlands
[*]Institute for Perception Research
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

ABSTRACT

The present research aims at a partial model for synthesized text intonation, restricted to utterances consisting of a single intonation phrase. Intonational characteristics of short utterances, embedded in systematically varied positions of a paragraph were investigated.

The effects of paragraph position on sentence prosody were mainly located in the beginning of the utterance. A model was formulated capturing the main findings and it was perceptually evaluated within the framework of a system for text-to-speech synthesis. The results indicated that the model contributed to the naturalness of the synthesized intonation.

10.21437/ICSLP.1992-97

I. INTRODUCTION

The number of publications in the area of text prosody is steadily growing. Descriptions are available for different languages, from different theoretical backgrounds [among others 1, 5, 6, 11, 12]. Research on the synthesis of intonation for text-to-speech conversion in Dutch has made clear that the overall intonation of a synthesized text does not yet sound very natural, due to the fact that most research has focused on the prosodic characteristics of sentences rather than texts. Therefore, there is a clear need for rules to generate text intonation for Dutch. Since there is no guarantee that text intonation is not language specific, we have to establish the Dutch phenomena from scratch.

Most of the existing models express text intonation by means of rules specifying the position of pitch minima and/or maxima in the time by frequency domain as a function of textual characteristics. The present research starts from an existing description of possible sentence melodies for Dutch by 't Hart, Collier and Cohen [3]. They account for the relations between pitch minima and maxima in terms of baselines and toplines. Until now, parameter specifications for baselines and toplines were obtained for isolated sentences only. We want to extend this existing model with parameter specifications for text intonation. We performed an experiment with read aloud speech. This enables us to keep text structure under experimental control and to study prosodic variation as a function of experimental manipulations. We take the written paragraph as a text unit. In the remainder of the paper we report acoustical measurements and formulate rules for generating text intonation. Also, a perceptual evaluation of the rules will be reported.

II. ANALYSIS OF PRODUCTION DATA

2.1 Introduction

In this section we report an analysis of the prosodic behavior of two speakers reading aloud several texts. In order to get a clear view of the effect of text structure on prosody, we present the speakers with different versions of the texts having the same utterance in different positions within a paragraph. In this way, the prosodic characteristics cannot be accounted for in terms of the content of the utterances, but can only be attributed to intersentential relations, i.e., to the position of the utterance in the paragraph.

2.2 Methods

We constructed ten texts of three paragraphs. Four versions of each text were made up. The first and the last paragraph were identical across versions. The second paragraph in each text contained five sentences. We used the second paragraph as the source of material to avoid text initial or text final effects. The four versions of each text differed with respect to the order of the sentences in the second paragraph. One of these sentences was used as the target sentence for the acoustical measurements. Across versions this sentence occurred in four relevant positions of this paragraph: 1, 2, 4 and 5 (the third position was not relevant for our aim [8]). Sometimes it was necessary to add some markers to the remaining sentences, to clarify the semantic relations within the paragraph. Of course, these markers never occurred in the target sentences themselves. The procedure is illustrated in the example below, where two translated instances of the same paragraph are presented with the target sentence in first and second position. The target sentence is underlined.

> Hay fever can seriously impair the capacity to think and concentrate. A severe attack of hay fever can put someone completely out of action. This can have dramatic consequences for some school children. It generally begins in early puberty and is at its most severe in the first few years. On top of that, the exams take place in the pollen season.
>
> A severe attack of hay fever can put someone completely out of action. Hay fever can seriously impair the capacity to think and concentrate.This can have dramatic consequences for some school children. It generally begins in early puberty and is at its most severe in the first few years. On top of that, the exams take place in the pollen season.

The texts were read aloud by two male native speakers of Dutch. They also read each target sentence in isolation. The recordings were made digitally in a sound isolated room.

The 20 (2 speakers * 10 sentences) isolated sentences and the 80 (2 speakers * 40 sentences) target sentences were isolated from their context and digitized (16kHz, 12 bits, 7.8 kHz LP) on a VAX/VMS computer.

A perception experiment was designed to investigate if listeners are able to judge what position the sentences, cut out from their context, had occupied in the paragraph. The conclusion from this experiment was that listeners indeed heard differences between the intonation of sentences in different positions (for more details we refer to [8]). On the basis of the results we selected the eight sentences whose original paragraph position most easily could be

recognized (≥75% correct) for subsequent acoustical analysis.

2.3 Measurements

The speech material selected was analyzed with a frame length of 10 ms, and 28 LPC coefficients. A pitch contour for each sentence was obtained using a pitch detection algorithm. Pitch tracking errors were corrected manually. Subsequently the pitch contours were stylized: the original craggy pitch contour is replaced by a stylized contour which is perceptually identical to the original contour. Perceptually irrelevant details are removed in this way and relevant pitch movements are preserved. The first part of the stylization was done automatically using a stylization program. This stylization is only based on properties of the signal. However, these stylizations sometimes leave irrelevant movements. Therefore, all the stylizations were corrected manually by two phonetically trained listeners, mostly on auditory criteria.

Three types of measurement were performed on the production data. The pitch maximum in the first accented syllable in each sentence was determined. In all cases the pitch maximum was part of an accent-lending pitch rise. The maximum was determined using visual criteria.

Furthermore, the course of the baseline (the line that optimally describes the valleys of the pitch contours) and the topline (running through the pitch peaks) were determined and described in terms of the onset and offset pitch of these lines (pitch was plotted on a logarithmic scale). In order to determine onsets and offsets their course was extrapolated to the first and the last sample of the sentence. We are aware of the problem, occurring with fitting baselines and toplines. However, it can be done in a much more reliable way than suggested by Lieberman [7] (i.e. as described by Huber [4]). Comparisons among paragraph positions were only made if the sentence melodies on the four positions were identical in terms of intonation transcription.

2.4 Results and discussion

A two-way analysis of variance was performed on the data with paragraph position as a fixed factor and speaker as a random factor. F0 of the first accent, onset frequency of the topline, offset frequency of the topline and onset and offset frequency of the baseline were used as dependent variables. There were no significant interactions between speaker and position (all cases: $F<1$). Therefore we decided to collapse the results for both speakers in our analyses. We performed one-way analyses of variance on the data with position as a fixed effect. First pitch peak, onset and offset of the topline and onset and offset of the baseline were used as dependent variables. A Student Newman-Keuls post hoc analysis was used to make pairwise comparisons between means. The results will be discussed below in separate subsections.

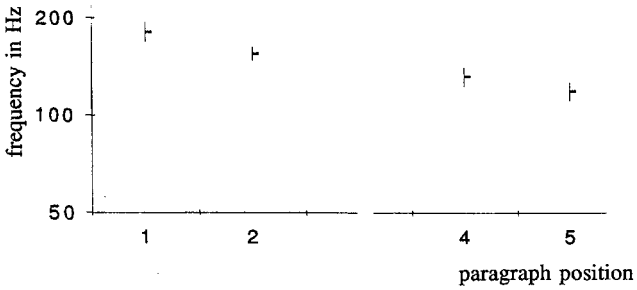

Fig. 1 Means and standard deviations (in Hz) of the first pitch maximum in the utterance as a function of utterance position within the paragraph

The first pitch maximum. In figure 1, means and standard deviations of the first pitch maximum are shown for each position. Each cell represents eight measurements.

Paragraph position is responsible for differences in the pitch of the first peak, starting high at the beginning of a paragraph and decreasing over the course of the paragraph. The overall effect for position was significant ($F_{(4,28)}=16.5$; $p<.001$). In addition, the first pitch maxima in isolated utterances are almost equal to the pitch maxima in paragraph-initial position. Finally, the data in fig. 1 suggest that the speaker's behaviour exhibits some form of supra-declination: there is a sequential lowering of the first pitch maximum of an utterance throughout the paragraph.

The topline and the baseline. In figure 2 the course of the baseline is shown as a function of the position in the paragraph. Offsets of the topline and the baseline are almost constant for the

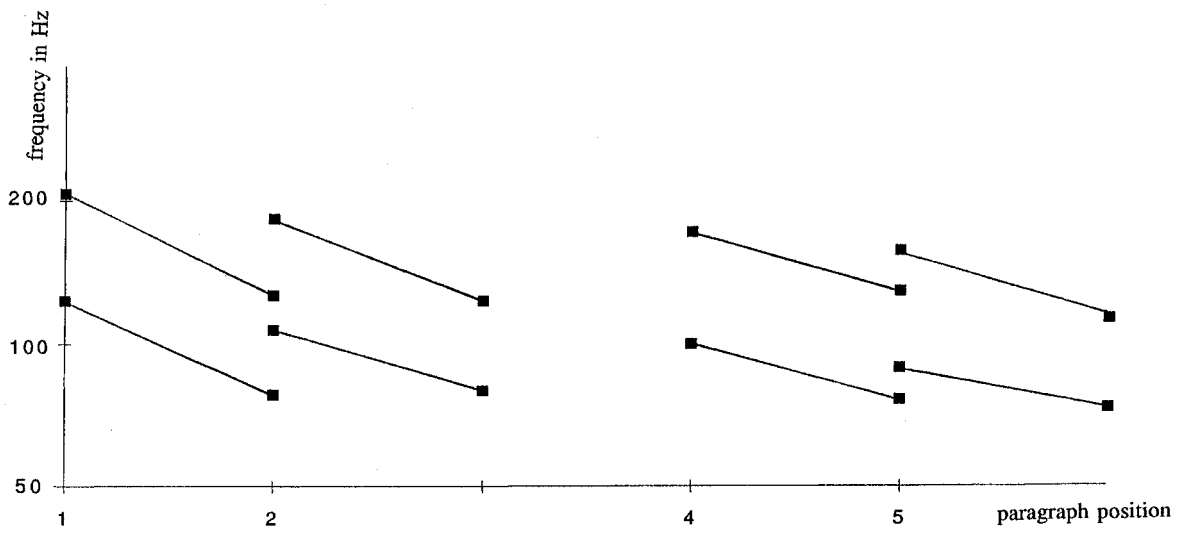

Fig. 2 Means (in Hz) of the onsets and offsets in the utterance as a function of utterance position within the paragraph

different paragraph positions (the differences are not significant: both cases $F_{(4,28)}<1$). Overall means for topline and baseline offset are 124 Hz and 78 Hz, respectively.

The overall effect of position on **onset frequency of the topline** is significant ($F_{(4,28)}=6.8$, $p<.001$). Post hoc comparisons show that we should distinguish the following three subsets: 1; 2 and 4; 4 and 5. This supports an interpretation in terms of sequential lowering of the onset value over the paragraph in line with the interpretation of the data for first peak. Isolated sentences have the same value as paragraph initial sentences.

The values of the **onset frequency of the baseline** also show an overall significant result ($F_{(4,28)}=5.4$, $p=.002$). Post hoc comparisons reveal the same subset structure as for topline onsets. These data also support the interpretation mentioned above. The overall conclusion is that both speakers show supra-declination in reading aloud texts. Isolated sentences have the same value as the medial positions.

We also calculated the distance between topline and baseline in semitones. The results are presented in table 1. We performed a oneway analysis of variance on the results with position as fixed factor.

Table 1. The distance between the topline and the baseline in semitones for the different paragraph positions (1-5) and for isolated utterances (ISO). Standard deviations are presented in italics.

position	1	2	4	5	ISO
onset	8.9	9.3	9.3	9.8	11.0
sd	*2.6*	*2.2*	*2.5*	*3.9*	*2.9*
offset	8.3	7.7	9.2	7.6	6.5
sd	*3.0*	*3.1*	*3.0*	*2.6*	*1.1*

There is no effect of position in the paragraph on the frequency interval between topline and baseline, neither for onset nor for offset frequency (both cases: $F<1$). Each sentence starts with a range of about 9.5 semitones and ends with a range of about 8 semitones. The sentence in isolation has a wider range at the beginning and a somewhat smaller range at the end.

III. RULES FOR TEXT INTONATION

The above results support the conclusion that speakers produce some kind of supra-declination throughout the paragraph. However, our aim is not to formulate a production model but to formulate a simple algorithm to synthesize intonation for automatic text-to-speech conversion. Therefore, perceptual tests were conducted with the target utterances taken from the production study to determine which productional distinctions are perceptually relevant. The results revealed that listeners were able to identify the initial and the final position, but they were not able to differentiate between the medial positions. We therefore formulate a perception based model allowing only three categories: initial position, medial position and final position.

The rules specify the following topline and baseline offsets and onsets in Hz for the three paragraph positions. The values are presented in the scheme below:

		initial	medial	final
Topline	onset	205 Hz	180 Hz	160 Hz
	offset	125 Hz	125 Hz	125 Hz
Baseline	onset	125 Hz	105 Hz	90 Hz
	offset	80 Hz	80 Hz	80 Hz

As can be seen, the topline and baseline offsets are the same for all positions. A perceptual evaluation of these rules is reported in the following section.

IV. PERCEPTUAL EVALUATION

4.1 Introduction

The aim of the perceptual evaluation is to investigate if listeners appreciate texts provided with the rules for paragraph intonation formulated in the previous section. The rules should contribute to a more natural prosody for coherent text in semi-automatic text-to-speech conversion. Therefore, the rules were incorporated within an existing algorithm for synthesizing intonation as part of a system for automatic text-to-speech conversion. The existing algorithm generates natural intonation contours for isolated utterances, by means of rules for declination resets and accent downstep [10].

4.2 Methods

Materials. We used a text of 95 words consisting of two paragraphs, each containing three sentences. Each paragraph expressed a separate news item. Table 2 summarizes the experimental conditions[2].

Table 2. Survey of experimental conditions.

	paragraph intonation	
	yes	*no*
decl. resets and accent downstep	*version 1*	*version 2*

The different versions were produced by generating the appropriate pitch contours on diphone speech (LPC 30, 20kHz, 16 bits [2]).

Subjects and procedure. 25 subjects participated in the evaluation. They rated the naturalness of all versions on a 10-point scale ("1" corresponding to "very unnatural", "10" corresponding to "very natural"). Each version was presented five times. The first presentation of each version was considered as a practice trial. Four different orders of presentation were used for different groups of subjects. Stimuli were presented over headphones.

4.3 Results and discussion

The original ratings were used as raw data for further analyses. Means and standard deviations, computed over subjects and replications, are shown in table 3.

Table 3. Means and standard deviations for each experimental condition.

	paragraph intonation	
	yes	*no*
decl. resets and accent downstep	***7.2*** $\pm$*1.8*	***6.3*** $\pm$*2.1*

As can be seen, there is an effect of 'presence or absence of paragraph intonation (7.2 vs. 6.3 on the naturalness scale, respectively). A oneway analysis of variance was performed on the data, with 'presence or absence of paragraph intonation' as a fixed factor. The results show that the presence of paragraph intonation gives a considerable increase in judged naturalness.

V. GENERAL CONCLUSION

The present results show that speakers use prosody to convey information about the global structure of a text. The detailed findings support observations for other languages, such as Danish

and Swedish [1, 11, 12], that speakers display some kind of supra-declination throughout a discourse segment in read text.

On the perceptual side, it was found that listeners appear to be insensitive to differences between the prosodic characteristics of paragraph-medial utterances produced by speakers. As a result, a simple set of rules for paragraph intonation could be formulated on the basis of the acoustic measurements. An evaluation of these rules showed that paragraph intonation is appreciated.

References

[1] G. Bruce. "Textual aspects of prosody in Swedish," *Phonetica*, vol. 39, pp. 274-287, 1982.

[2] R. Drullman and R. Collier. "Speech synthesis with accented and unaccented diphones," in: V.J. van Heuven and L.C.W. Pols, eds., *Analysis and synthesis of speech, strategic research towards high-quality text-to-speech generation.* Berlin: Mouton de Gruyter, 1992 (in press).

[3] J. 't Hart, R. Collier and A. Cohen. *A perceptual study of intonation.* Cambridge: University Press. 1990.

[4] D. Huber. "A statistical approach to the segmentation and broad classification of continuous speech into phrase-sized information units" Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-89) Glasgow, 1989.

[5] D.R. Ladd. "Declination reset and the hierarchical organization of utterances," *J. Acoust. Soc. Am.*, vol. 84, pp. 530-544, 1988.

[6] Lehiste, I. "Phonetic characteristics of discourse," Acoust. Soc. Jpn., Trans. Committee Speech Res. (April 1980), pp. 25-38. 1980.

[7] P. Lieberman, W. Katz, A. Jongman, R. Zimmerman and M. Miller "Measures of the sentence intonation of read and spontaneous speech in American English," *J. Acoust. Soc. Am*, vol. 77, pp. 649-657, 1985.

[8] A.M.C. Sluijter. [Text intonation, an acoustical and perceptual study of the relation between text structure and F0 course, (in Dutch)], Report no. 774, Inst. for Perception Research, Eindhoven, 1991.

[9] A.M.C. Sluijter. [Perceptual evaluation of a model for paragraph intonation with synthetic speech, (in Dutch)], Report no. 801, Inst. for Perception Research, Eindhoven, 1991.

[10] J.M.B. Terken, Synthesizing natural sounding intonation for Dutch: rules and perceptual evaluation, Manuscript no. 769, Inst. for Perception Research, Eindhoven, 1990.

[11] N. Thorsen, "Intonation and Text in Standard Danish" in: *J. Acoust. Soc. Am.*, vol 77, pp. 1205-1216, 1985

[12] N. Thorsen "Sentence intonation in textual context, Supplementary data" In: annual report of the institute of Phonetics, University of Copenhagen 20, pp. 35-44, 1986.

Notes

1. This research was done at the Institute for Perception Research, Eindhoven as part of the first author's MA-thesis, which was written under the supervision of the second author. We thank Hugo Quené and Vincent van Heuven for ideas and discussion.

2. The original experiment comprised more experimental conditions. The other versions of the text did not reach scores above 5.4 on the naturalness scale. For more details see ref. [9].